PRESENTATION ON

MUSIC CLASSIFICATION

Using 2D Mel-spectrogram and CNN

Aadesh Dhakal

MTechAI 2023

Kathmandu University

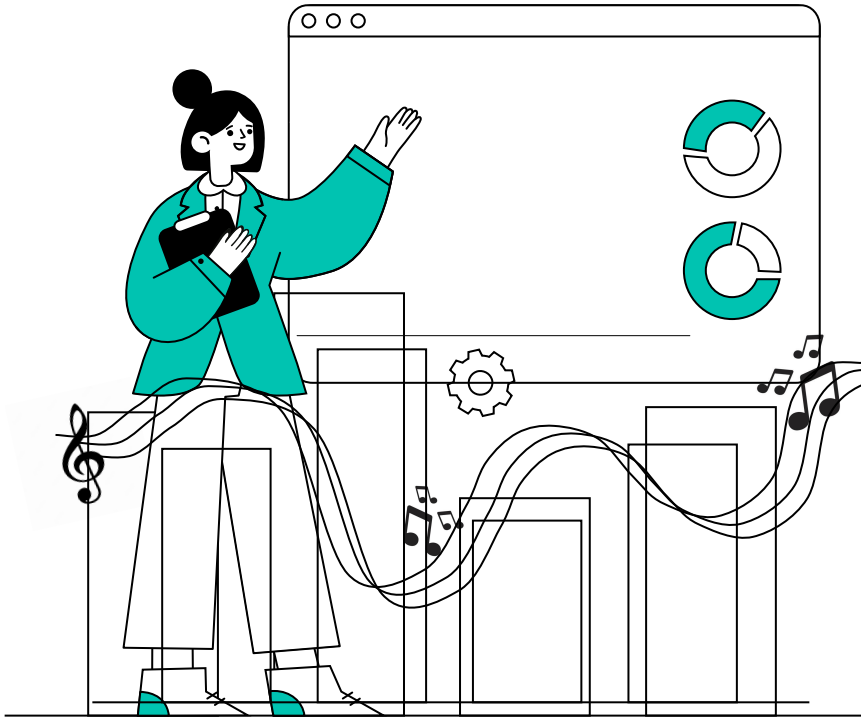
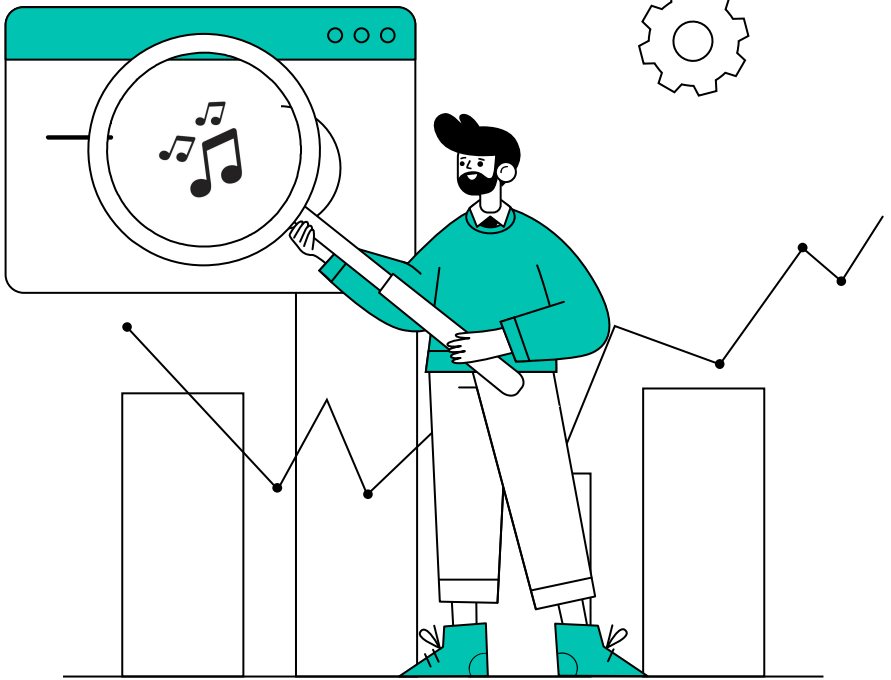


TABLE OF CONTENTS

PROBLEM STATEMENT

Capturing information from audio using 2D Mel-Spectrogram and classifying music genre using Convolutional Neural Network

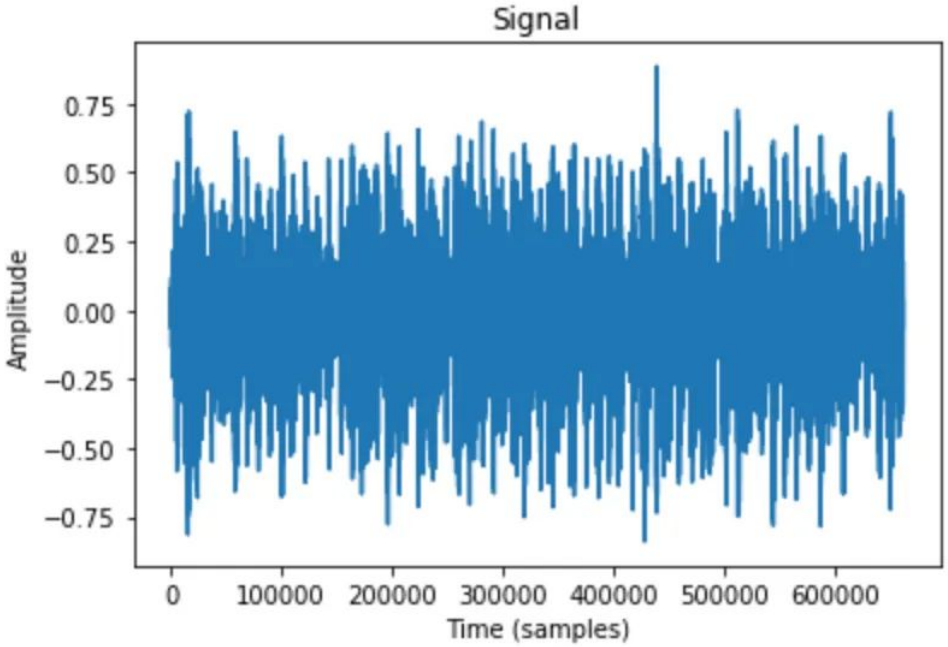
01 CAPTURING AUDIO



CAPTURING AUDIO

- ➔ A signal is a variation in certain quantity over time.
- ➔ For audio, the quantity that varies is air pressure.
- ➔ Capturing audio data is basically **taking samples of air pressure over time.**
- ➔ Common used rate is 44.1kHz, and this captured entity is called a waveform of a signal.

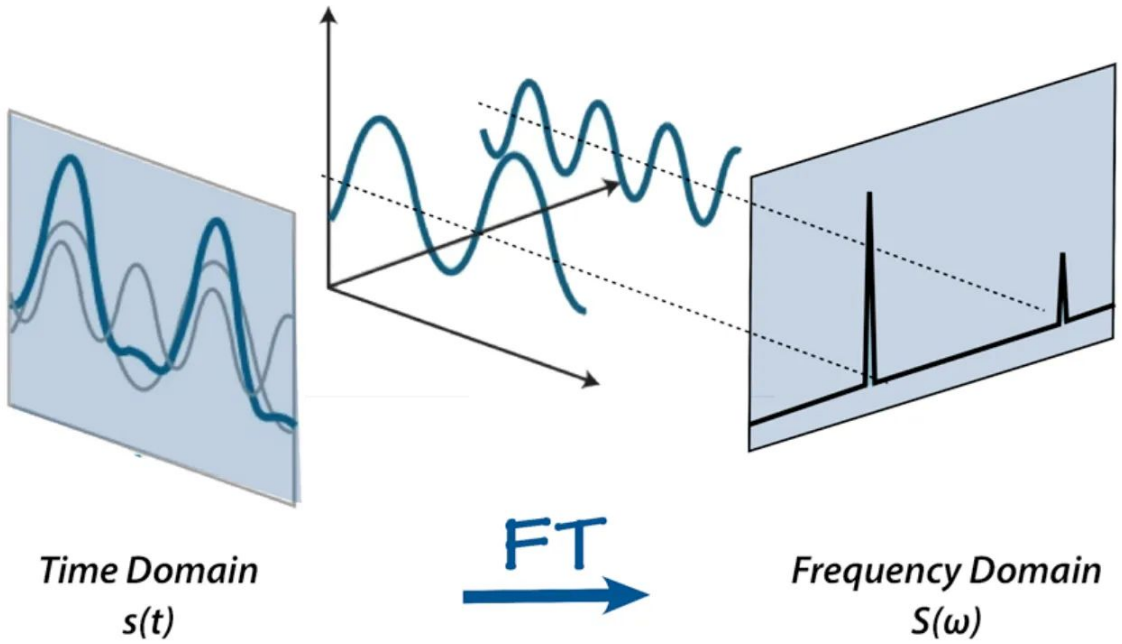
A WAVEFORM

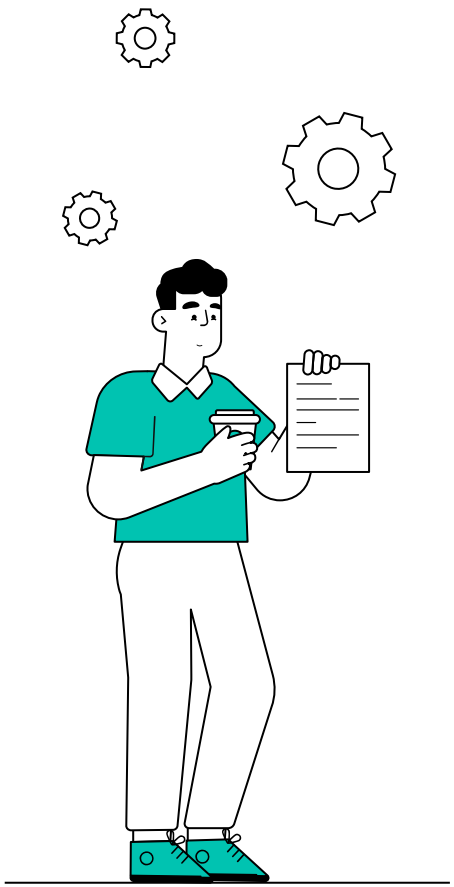

Signal
Amplitude
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
100000
200000
300000
400000
500000
600000
Time (samples)

HOW DO WE CAPTURE INFORMATION FROM A WAVEFORM?

- ➔ The answer is **Fourier.**
- ➔ An audio signal is comprised of several single-frequency sound waves.
- ➔ When taking samples, we are only capturing amplitude over time.
- ➔ **Fourier transformation** allows decomposition of a signal into individual frequencies and its amplitude.
- ➔ **The Fast Fourier Transform (FFT)** is an algorithm that can efficiently compute the Fourier transform.

FOURIER TRANSFORMATION





02

FOURIER'S THEOREM

FOURIER'S THEOREM

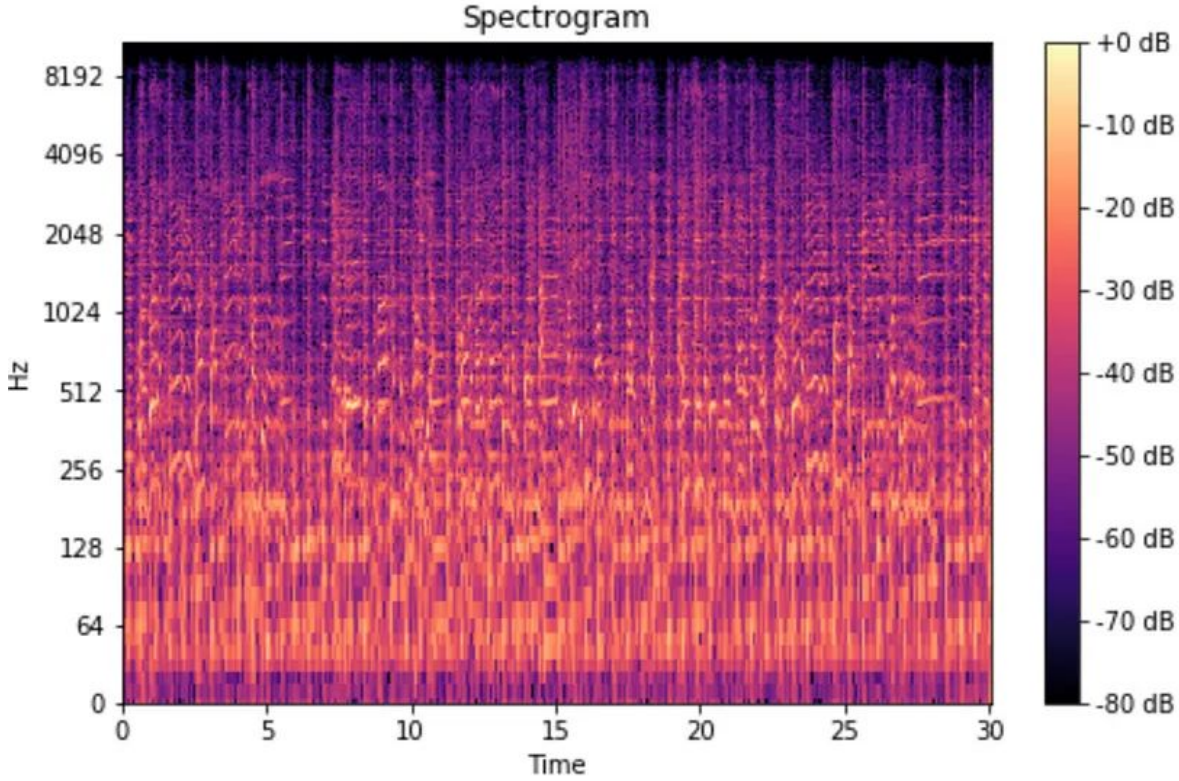
- ➔ Every signal can be decomposed into a **set of sine and cosine waves** that add up to the original signal.
- ➔ Fourier's transformation allows us to analyze the frequency content of a signal. But for **non-periodic signals**, we need **spectrogram** of such signals.
- ➔ This is called the **short-time Fourier Transform.**

FOURIER'S THEOREM

- The FFT is computed on overlapping windowed segments of the signal, and the result is a spectrogram.
- Basically, a spectrogram is a group of FFts in a stack.
- It is a way to visually represent a signal's amplitude as it varies over time at different frequencies..

03 MEL-SPECTROGRAM

THE SPECTROGRAM



WHY DO WE NEED MEL-SPECTROGRAM?

- ➔ Humans perceive frequency **logarithmically.** (Not on a linear scale)
- ➔ Humans are better at detecting differences in lower frequencies than higher frequencies.
- ➔ A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. (mel for melody)
- ➔ An Experiment **link**

AGAIN, WHY MEL-SPECTROGRAM?

Ideal audio feature must include:

1) Time frequency representation
2) Perceptually relevant amplitude representation
3) Perceptually relevant frequency representation (Mel-spectrogram)

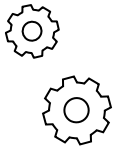
$$mel = 1127.01048 * \log(f/700 + 1)$$

2D MEL-SPECTROGRAM?

- ➔ Audio data is represented in a **two dimensional array.**
- ➔ One represents **time (x-axis)**, and the other **frequency (y-axis).**
- ➔ Each element in the array corresponds to the energy (power) of a specific frequency bin at a specific time.

04

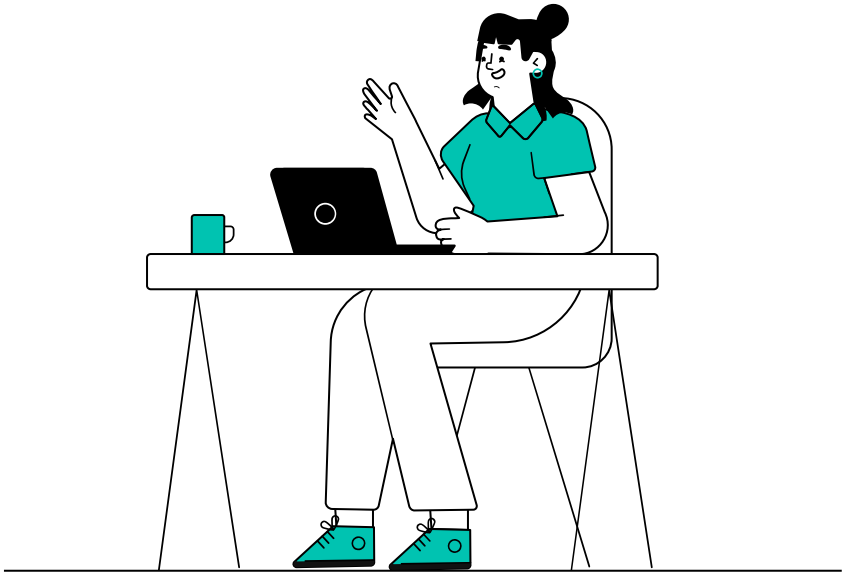
DEMONSTRATION



```
100/100 [==============================] - 4s 38ms/step - loss: 1.0632 - accuracy: 0.7760 - val_loss: 1.0762 - val_accuracy: 0.7950
Epoch 73/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0934 - accuracy: 0.7447 - val_loss: 1.0647 - val_accuracy: 0.8050
Epoch 74/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0726 - accuracy: 0.7610 - val_loss: 1.0813 - val_accuracy: 0.7700
Epoch 75/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0675 - accuracy: 0.7710 - val_loss: 1.1027 - val_accuracy: 0.7750
Epoch 76/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0918 - accuracy: 0.7384 - val_loss: 1.0587 - val_accuracy: 0.8000
Epoch 77/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0525 - accuracy: 0.7622 - val_loss: 1.0691 - val_accuracy: 0.7850
Epoch 78/80
100/100 [==============================] - 4s 39ms/step - loss: 1.0635 - accuracy: 0.7685 - val_loss: 1.0827 - val_accuracy: 0.7650
Epoch 79/80
100/100 [==============================] - 4s 39ms/step - loss: 1.0622 - accuracy: 0.7685 - val_loss: 1.1364 - val_accuracy: 0.7500
Epoch 80/80
100/100 [==============================] - 4s 38ms/step - loss: 1.0414 - accuracy: 0.7710 - val_loss: 1.0816 - val_accuracy: 0.7900
2024-05-27 02:18:56.683852: W tensorflow/core/common_runtime/bfc_allocator.cc:290] Allocator (GPU_0_bfc) ran out of memory trying to allocate 788.17MiB with freed_by_count
=0. The caller indicates that this is not a failure, but this may mean that there could be performance gains if more memory were available.
2024-05-27 02:18:56.684180: W tensorflow/core/common_runtime/bfc_allocator.cc:290] Allocator (GPU_0_bfc) ran out of memory trying to allocate 788.17MiB with freed_by_count
=0. The caller indicates that this is not a failure, but this may mean that there could be performance gains if more memory were available.
7/7 [==============================] - 1s 29ms/step - loss: 1.0816 - accuracy: 0.7900

(myenv) C:\Users\asus\OneDrive\Desktop\py\deeplearning>
```

05 Q / A AND CONCLUSION



REFERENCES

- https://www.youtube.com/watch?v=UKHBWzoOKsY (Fourier)
- https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53 (Mel-Spectrograms)
- https://www.youtube.com/watch?v=9GHCiiDLHQ4
- https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0
- https://www.kaggle.com/code/andradaolteanu/work-w-audio-data-visualise-classify-recommend/notebook (DataSet)

THANK YOU